SHOULD (T_1- T_2) HAVE LARGER UNCERTAINTY THAN T_1?

William F. Guthrie

National Institute of Standards and Technology, Gaithersburg, MD, USA

ABSTRACT

In interlaboratory comparisons, laboratories sometimes use a transfer instrument to realize the value of a laboratory standard to compare the relative biases of their measurement processes and standards. One summary of interest from such comparisons is the pairwise difference between two laboratories' results, along with its expanded uncertainty, a confidence interval for the true difference. Since the labs have unequal variances, the confidence interval is usually computed by the Welch-Satterthwaite (denoted WS) procedure, which approximates the distribution of the pivot quantity used to compute the confidence interval by a Student's-*t* distribution with effective degrees of freedom defined as a function of the data.

In the course of analyzing the data from a comparison of temperature realizations, an awkward and counterintuitive property of the WS procedure was observed. Namely, a confidence interval for a between-lab difference can be narrower than the corresponding interval for one of the component results. This occurs when at least one laboratory's uncertainty estimate has low degrees of freedom (say 1 or 2), and therefore has a large coverage factor from the Student's-*t* distribution, while the effective degrees of freedom for the combined uncertainty of the pairwise difference, obtained from the WS approximation, is larger.

The typical reaction to this situation is to suspect the WS procedure of failing to achieve its nominal confidence level. However, this is not the correct explanation. In fact, situations exist where the confidence intervals for each laboratory's mean and for their pairwise difference all achieve the stated level of confidence even though the uncertainty of the difference is smaller than the uncertainty of at least one of its component results. This paper explains how this counterintuitive property of confidence intervals can be true.

1. INTRODUCTION

In interlaboratory comparisons, two or more laboratories sometimes use a transfer instrument to realize the value of a laboratory standard to compare the relative biases of their measurement processes and standards. One summary of interest from such comparisons is an estimate of the pairwise difference, T_1-T_2, between two laboratories' mean results, along with an expanded uncertainty for the estimated difference. To aid in its interpretation, the expanded uncertainty should capture the true unknown difference between the laboratories' measurement results with a high probability, prespecified by the users of the comparison. Since the labs cannot reasonably be assumed to have the same levels of uncertainty in their measurement results, this confidence interval for the true difference between the laboratories' results is usually computed using the Welch-Satterthwaite (denoted WS) procedure [1,2].

The WS procedure approximates the distribution of the pivot quantity, *T*,

$$T = [(T_1\text{-}T_2)\text{-}(m_1\text{-}m_2)]/(u_1^2+u_2^2)^{1/2} \tag{1}$$

used for computation of a confidence interval with a Student's-*t* distribution with effective degrees of freedom, $v_{1\text{-}2}$, defined as

$$v_{1\text{-}2} = [(u_1^2+u_2^2)^2]/[(u_1^4/v_1)+ (u_2^4/v_2)] \tag{2}$$

where T_i is the observed mean measurement result for the i^{th} laboratory, m_i is its unknown, true mean temperature measurement result (as if obtained from an infinite number of replicate measurements), u_i

is the estimated combined standard uncertainty of T_i, and v_i is the degrees of freedom for u_i. Use of an approximation like the WS is necessary because no exact distributional theory currently exists for this type of statistic, a linear combination of normally-distributed quantities with unequal variances.

Although it is only an approximation to the unknown distribution of T, the WS procedure is known to work well in many practical situations. Recently, however, in the course of analyzing data from a comparison of temperature realizations, an awkward and counterintuitive property of the WS procedure was observed. Namely, the confidence interval for a between-lab difference, m_1-m_2, can be narrower than one or both of the corresponding intervals for the component results, m_1 or m_2.

For example, suppose one laboratory calibrated an SPRT at a particular fixed point using the mean of the results from two separate freezes. The SPRT is then sent to another laboratory, which calibrates the thermometer using four freezes. With the resulting uncertainty budgets shown in Table 1, below, the expanded uncertainty for the difference of the two laboratories will be less than the expanded uncertainty for the first laboratory's calibration, at the 95% confidence level.

Table 1: Uncertainties for two laboratories' calibrations of a transfer SPRT at the Zn Point and the uncertainty of the difference between the laboratories' temperature realizations.

	Lab 1		**Lab 2**		**Lab 1 - Lab 2**	
Source/Type of Uncertainty	u_{1i}	v_{1i}	u_{2i}	v_{2i}	$u_{(1\text{-}2)i}$	$v_{(1\text{-}2)i}$
Freeze-to-Freeze Repeatability (A)	0.579	1.000	0.393	3.000	0.700	1.993
Hydrostatic Head (B)	0.003	∞	0.017	∞	0.017	∞
Chemical Impurities (B)	0.121	∞	0.143	∞	0.187	∞
Heat Flux - Immersion (B)	0.100	∞	0.100	∞	0.141	∞
Nonlinearity of Bridge (B)	0.029	∞	0.081	∞	0.086	∞
Stability of Reference Resistor (B)	0.005	∞	0.008	∞	0.009	∞
Propagated TPW (B)	0.169	∞	0.130	∞	0.213	∞
SPRT Self-Heating (B)	0.058	∞	0.031	∞	0.066	∞
Pressure (B)	0.004	∞	0.043	∞	0.043	∞
Plateau (B)	0.078	∞	0.060	∞	0.098	∞
Combined Uncertainty ($u_1, u_2, u_{1\text{-}2}$**)**	**0.631**		**0.464**		**0.784**	
Degrees of Freedom (v_1, v_2, $v_{1\text{-}2}$**)**	**1.415**		**5.820**		**3.132**	
Coverage Factor ($k_1, k_2, k_{1\text{-}2}$**)**	**6.550**		**2.465**		**3.108**	
Expanded Uncertainty ($U_1, U_2, U_{1\text{-}2}$**)**	**4.136**		**1.143**		**2.435**	

This property of the WS approximation has also been noticed and reported on by Ballico [3] and Hall and Willink [4] recently, who discuss other examples of this phenomenon, some temperature-related.

A typical, and not at all unreasonable, reaction to anecdotal examples of this phenomenon is to suspect the WS procedure of failing to achieve its nominally-specified confidence level. However, this is not the correct explanation. In fact, there are situations for which all of the confidence intervals involved in these examples achieve their stated confidence levels, yet the expanded uncertainties are ordered in contradiction to our expectations. This paper attempts to explain why this phenomenon occurs and how these counterintutitive results can be valid, using the example given in Table 1.

2. A HYPOTHESIS FOR THE COUNTERINTUITIVE ORDERING OF UNCERTAINTIES

The example presented in Table 1 gives some clues as to why this phenomenon occurs. The very low degrees of freedom for the estimate of freeze-to-freeze repeatability for Lab 1 are an unusual feature of this example, relative to the assumptions often made (implicitly or explicitly [1,2]) when thinking about the properties of uncertainty estimates. The fact that the repeatability is the largest uncertainty component means that the effective degrees of freedom of u_1 are also low. Because of that, the coverage factor, k_1, obtained from the Student's-*t* distribution, used to compute U_1 is very large. This yields a large expanded uncertainty for the calibration at Lab 1. In addition, the uncertainty for the calibration at Lab 2 is fairly similar in size to the uncertainty at Lab 1, though based on a larger number of measurements. Therefore, when the uncertainty of the difference in the temperature realizations is computed, its effective degrees of freedom are slightly higher than the effective degrees of freedom for the calibration at Lab 1. Because v_1 and v_{1-2} are both low, however, the difference between their coverage factors is large, while the standard uncertainty for the difference, u_{1-2}, is only slightly greater that u_1. As a result, U_1 and U_{1-2} end up having magnitudes ordered in reverse of our expectations. A key point, however, is that the problem seems to be caused by the coverage factors, k_1 and k_{1-2}, rather than the standard uncertainties, u_1 and u_{1-2}, which are ordered as expected.

The suggestion that the coverage factors are the immediate source of this problem indicates that the uncertainty in the uncertainty estimates u_1, u_2, and u_{1-2} are more important to this phenomenon, than the exact values of u_1, u_2, and u_{1-2} themselves. This is because the coverage factors obtained from Student's-*t* distribution are a direct measure of the additional uncertainty over and above u_1, u_2, or u_{1-2} required to compute valid confidence intervals because u_1, u_2, and u_{1-2} are estimates rather than being exactly known quantities. If the true values of the uncertainties associated with the estimates u_1, u_2, and u_{1-2} were available, then the values of k_1, k_2, and k_{1-2} would be obtained from the normal distribution, and would all be set to a value of 1.96 for a 95% confidence level. Instead, though, the values of the coverage factors range from roughly 2.5 to 6.6 to account for the variation inherent in u_1, u_2, and u_{1-2}. The effective degrees of freedom for u_1, u_2, and u_{1-2}, obtained directly from the WS formula, also tell us essentially the same thing that the coverage factors do, since there is a direct, one-to-one correspondence between the two quantities for any given confidence level.

Now, focusing on the uncertainty of u_1 and u_{1-2}, if some of the wildest variations in the estimates of u_1 could be balanced out by the associated estimates of u_2, then it would follow that repeated realizations of u_{1-2} actually should be less variable than repeated realizations of u_1. For example, sometimes when u_1 takes on a value that is lower than its true value, u_2 will take on a value higher its true value, allowing u_{1-2} to be closer to its true value more often then u_1 is. This would justify the behavior of U_1 and U_{1-2}, which summarize the total amounts of variation associated with the computation of confidence bounds. If u_{1-2} and u_1 are similar in size, but u_{1-2} has a little less uncertainty than u_1 does, then k_{1-2} should be less than k_1 and, consequently, U_{1-2} can be less than U_1 without indicating that the interval computed using the WS approximation has not attained its nominal confidence level.

3. VALIDATION OF THE HYPOTHESIS

Because the WS procedure is an approximation, mathematical proofs of hypotheses about its exact behavior are not likely to be easily obtained. Given that mathematical proof is not a very practical option, a good alternative for determination of its statistical properties is to directly observe the performance of the WS procedure under conditions where the true outcome of each result is known. Therefore simulation was used to test the hypothesis that $U_{1-2}<U_1$ does not imply failure of statistical intervals computed using the WS approximation because the uncertainty of u_{1-2} is less than the uncertainty of u_1. Accordingly, 100,000 sets of data were simulated from each of the two measurement processes outlined in Table 1 using random number generation. The normally-distributed estimates of the process means, m_1 and m_2, were each set to be 692.677 K. The estimates of u_1 and u_2 were generated to have scaled Chi-Squared distributions with the degrees of freedom set to 1.45 and 3.98, respectively. The true value of u_1 was set to 0.631 mK and the true value of u_2 was set to 0.439 mK.

Plots a_1-a_3 in the upper left of Figure 1 show the first 100 confidence intervals computed from the simulated data. The true values for each measurement process (offset by 692.677 K) and their true difference are denoted as solid horizontal lines. The plots show that most of the time the confidence intervals for each process cover their associated true parameter values. In fact, of the 100,000 intervals for m_1, 94,907 covered their true values, just as they should for 95% confidence intervals. Similarly the observed coverage probability for the confidence intervals for m_2 was 95.079%, also as expected. Finally, despite the fact that U_{1-2} was typically about 30% less than U_1 in 85,607 of the 100,000 sets of data (see, for example, the results of simulation 100), the observed coverage of m_1-m_2 was 94.873%. The fact that the observed coverage probabilities for all three sets of intervals are consistent with their nominal confidence level of 95% establishes this example as a counter-example to the (incorrect) conclusion that $U_{1-2}<U_1$ implies that the WS intervals are not valid confidence intervals. By definition the only criterion that a confidence interval needs to meet in order to be a valid confidence interval is the criterion that it attains its stated coverage probability.

In order to rest satisfied in the correctness of this counter-intuitive result, however, additional insight into its cause is helpful. Studying the individual simulation results in plots a_1-a_3 in Figure 1 sheds some light on the situation. A confidence interval may not capture its associated true value because all the measurements for that particular sample happened to be higher or lower than the true value by chance or because the measurements for a particular sample are closer together than they typically would be. For example, in simulation 40, the interval for m_1 (692678.20 mK ± 0.48 mK) does not cover its true value (692677 mK) because the two measurements from that sample happened to both be high and to

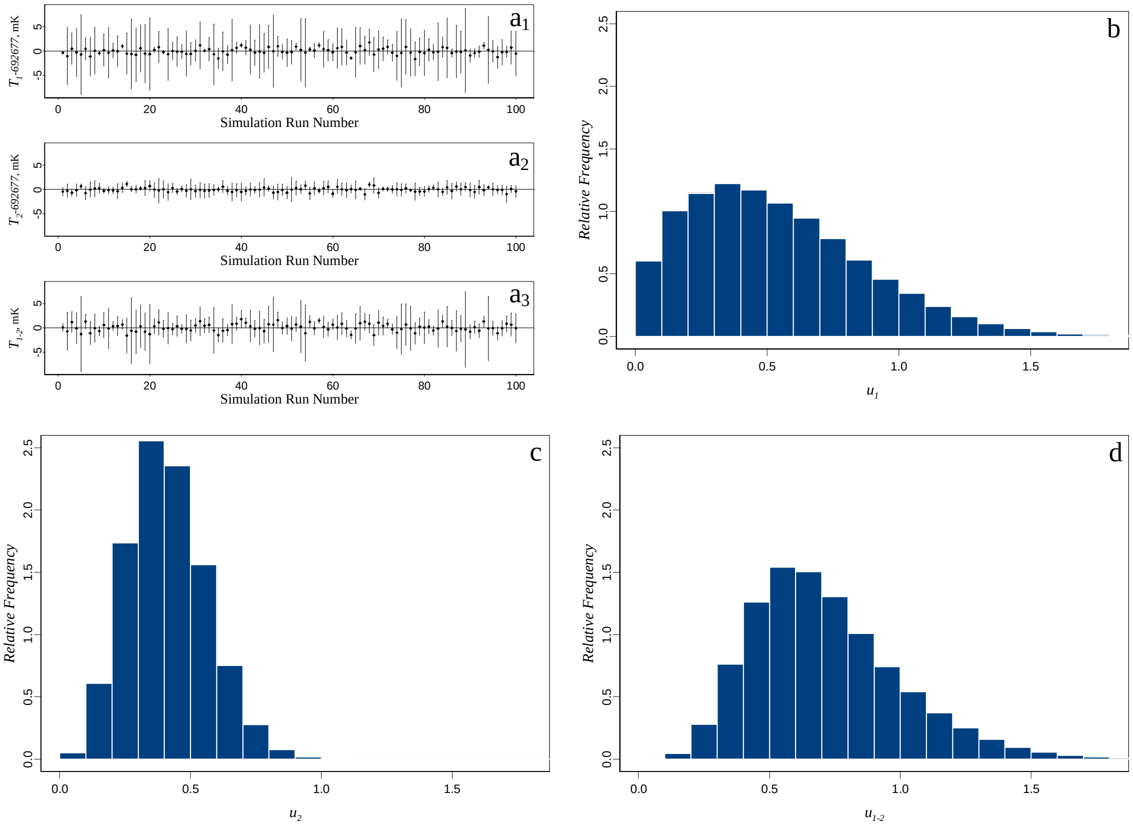

Figure 1: Plots showing the first 100 confidence intervals for m_1, m_2, and m_1-m_2 (plots a_1, a_2, and a_3) and histograms of all 100,000 values of u_1, u_2, and u_{1-2} (plots b, c, and d) obtained by simulation.

be very close together. The interval for m_2 (692676.46 mK ± 1.85 mK) does cover its true value, however, and also has a larger than typical estimate of u_2. As a result, although the interval for m_1-m_2 (692678.74 mK ± 1.84 mK) is centered relatively far away its true value, it covers its true value because the two uncertainty estimates balance out in the uncertainty estimate for the difference. This offers some insight into how the variation in uncertainty estimates might be reduced when individual uncertainty estimates are combined to estimate the uncertainty of the difference of two quantities.

Plots b, c, and d in Figure 1 provide statistical evidence that the uncertainty of u_{1-2} is less than the uncertainty of u_1, thereby explaining how U_{1-2} can be less than U_1 for valid confidence intervals. Plots b, c, and d are histograms of u_1, u_2, and u_{1-2} respectively. These histograms show how the distributions of u_1, u_2, and u_{1-2} compare, in particular with respect to their uncertainties. The fact that the values of u_{1-2} are more tightly clustered around their typical value than the values of u_1 are around theirs indicates that u_{1-2} is less variable than u_1. Similarly, comparing the lower tails of the distributions of u_1 and u_{1-2} shows that u_1 is much more likely to be underestimated than u_{1-2} is, while the upper tails of the two distributions are similar. Underestimation of u_{1-2} is less likely than underestimation of u_1 because u_2, a component of u_{1-2}, is estimated relatively precisely, bounding u_{1-2} below. This confirms the hypothesis that uncertainty in the uncertainty estimates accounts for the shortening of U_{1-2} relative to U_1. In fact, the histograms also show that the typical value of u_{1-2} is slightly greater than the typical value of u_1, as expected. However, the increase in u_{1-2} relative to u_1 is outweighed by the decrease in the coverage factors, k_{1-2} relative to k_1, which give the uncertainty in u_{1-2} and u_1.

Even though $U_{1-2}<U_1$ does not indicate a failure of the WS procedure to produce valid confidence intervals, that does not mean that the WS procedure always works perfectly. Being an approximation, there are some conditions under which it works better than others. Figure 2 shows additional results from simulations of the general scenario of Table 1 but with differing values of v_1 (1 and 3), v_2 (1 to 30 in steps of 0.5), and different ratios of u_2/u_1 (0.1 to 1 in steps of about 0.015). Again, 100,000 data sets were generated for each set of simulation conditions. Plots a and b show contours of equal attained coverage probability and equal mean values of U_{1-2}/U_1, respectively, for $v_1=1$. The fact that the attained coverage of intervals with a nominal confidence level of 95% drops as low as 88% for some conditions clearly shows that the WS procedure does not always work well. However, since the all of the contours of U_{1-2}/U_1 are less than one and 95% coverage is attained under some conditions, interval length cannot explain this failure of the WS procedure. Fortunately, plot c, for $v_1=3$, suggests that the WS procedure will approximately attain its stated coverage as long as the degrees of freedom for the dominant uncertainty component are not extremely low. Plot d, also for $v_1=3$, shows that under these conditions the lengths of the intervals will also be in much closer agreement with our expectations, though still not strictly ordered as intuition suggests they should be.

4. CONCLUSIONS

This paper illustrates a counterintuitive feature of confidence intervals computed by the WS procedure which is likely to be observed in interlaboratory comparisons when pairwise differences of results are used to compare measurement processes. However, this example is a counter-example to the intuitive response that this phenomenon indicates that the WS approximation fails to produce valid confidence intervals. Confidence intervals for each lab’s true measurement process mean and their true difference all attain their stated confidence levels even though the lengths of the intervals are not ordered as intuitively expected. This phenomenon arises because confidence limits must account for the uncertainty in both the observed measurement values and in the uncertainty estimates themselves. Although relative interval length is not a indicator of confidence interval validity, this fact does not mean that the WS procedure is without flaws. Problems with interval coverage occur when uncertainty estimates being combined differ in magnitude and precision.

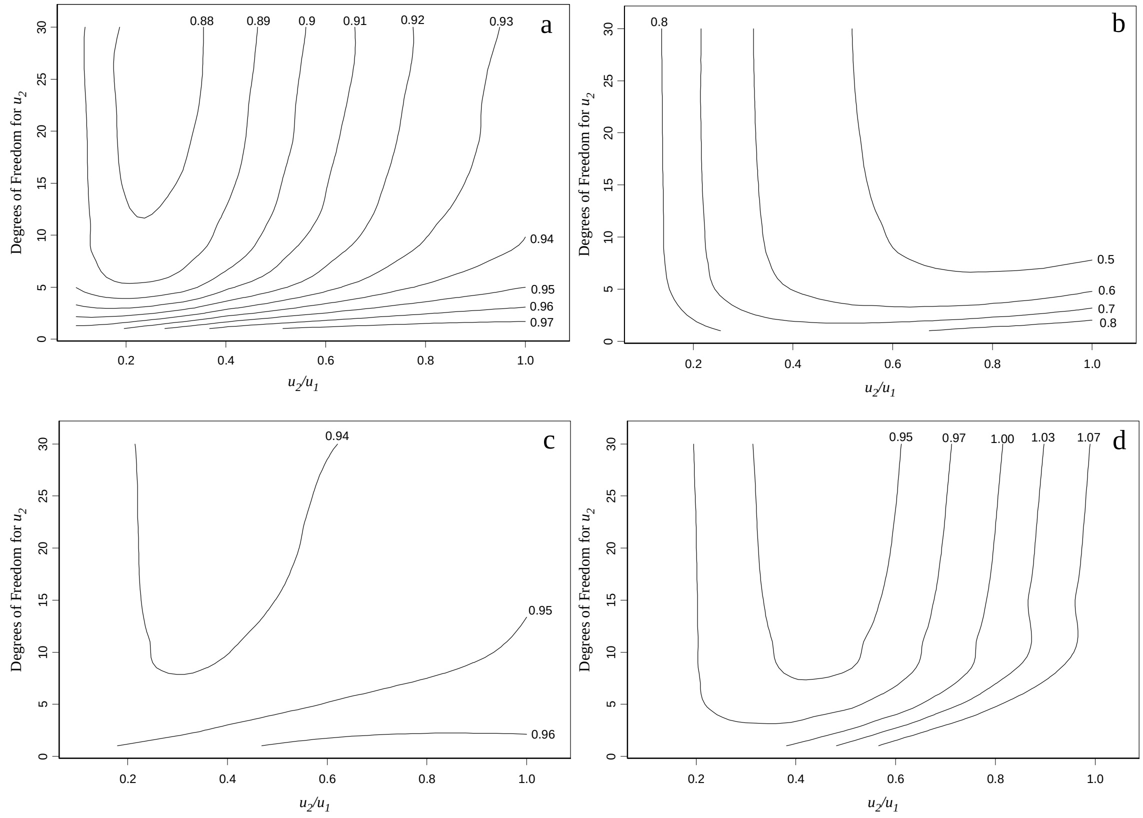

Figure 2: Contour plots showing the coverage probability attained for different values of v_2 and u_2/u_1 when v_1=1 and v_1=3 (plots a and c, respectively) and corresponding plots of U_{1-2}/U_1 (plots b and d,).

ACKNOWLEDGEMENTS

I would like to thank my colleagues in statistics and temperature metrology for their interest in this work. Keith Eberhardt, former Chief of the NIST Statistical Engineering Division, deserves special thanks for his willingness to discuss the statistical aspects of this project and encourage its pursuit.

REFERENCES

1. *Guide to the Expression of Uncertainty in Measuremen*t, Geneva, International Organization for Standardization, 1993, ISBN 92-67-10188-9.
2. Taylor, B.N. and Kuyatt, C.E., *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Technical Note 1297, 1994.
3. Ballico M., *Metrologia*, 2000, **37**, 61-64.
4. Hall, B.D. and Willink, R., *Metrologia*, 2001, **38**, 9-15.

Address of the Author:

William F. Guthrie, National Institute of Standards and Technology, Statistical Engineering Division, 100 Bureau Drive, Stop 8980, Gaithersburg, MD 20899-8980, USA, E-mail: will.guthrie@nist.gov, Internet: http://www.nist.gov/itl/div898/